ISSN 0317-2473

Research Discussion Paper No. 86

DETERMINATION OF SAMPLE SIZE FOR SURVEYS

by

N.M. Lalu

Population Research Laboratory
Department of Sociology
University of Alberta

NOVEMBER 1991

92-04420/

Determination of Sample Size for Surveys

Introduction

In the planning of a sample survey a stage is always reached when some decision must be made about the size of the sample. Most researchers have limited resources and so this decision is an important one. If the sample size is too large, the resources are wasted and a very small sample reduces the utility of the results. There are several ways of approaching the problem. Some people take an empirical view and seek the sample sizes that have been used by others in similar situations (Rossi et. al., 1983). This approach can be useful to inexperienced researchers. Sudman (1974) reported that the sample sizes of most national surveys are around 1000 or more and most of the regional surveys use sample sizes between 500 and 1000. In this paper a more formal approach is taken. The theories in sampling and probability have been used by statisticians to find a formal solution to the problem of determining the sample size. In this paper a brief summary of this formal approach and three computer programmes to solve the equations will be presented. The copies of the programmes may be obtained from the author. Sample outputs from the programmes are also given .

Sample Size required for the Estimation of Proportions to a Desired Degree of Accuracy.

A hypothetical example will illustrate the problem. Suppose a social scientist is interested in estimating the percentage of smokers among the males in a community. In this case one has to decide to what degree of accuracy the percentage is to be estimated. Suppose the researcher is satisfied to get the percentage within ± 5 percent. Even with this specification one cannot absolutely guarantee the accuracy within 5 percent. However large a sample one may take it is possible to get an unlucky sample which is in error by more than the desired five percent. Using sampling theory it is possible to control the chance of getting these unlucky samples. Suppose the researcher is willing to take a one in 20 chance of getting an unlucky sample where the error is more than five percent. With these specifications we are in a position to make an estimate of the required sample size provided the population size of the community is very large. We may safely say that the population is large if it is more than 20,000.

In technical terms, p, the estimate from the sample should be in the range of $P \pm 0.05$ except for a one in 20 chance, where P is the unknown proportion of smokers in the population. Since p is normally distributed about P, it will lie in the range $P \pm 1.96\sigma_p$

with a probability of 19 out of 20. σ_p is approximately equal to $\sqrt{\frac{PQ}{n}}$ where Q is equal to (1-P). Since the required error range of p is .05 we can write,

$$.05 = 1.96 \sqrt{\frac{PQ}{n}} \qquad (1)$$

Hence we may solve for n from the above equation in terms of P as

$$n = \frac{1.96^2\ PQ}{.0025} \qquad (2)$$

At this juncture the main problem is that the sample size depends on the same property of the population which we wish to estimate. If we have some idea of the value of P the sample can be calculated from equation 2. Now if we assume that P is around .5 the required sample size is 384. Any other assumption about P will result in a smaller sample size. For example, if we put P as .3 in equation 2 the sample size required is 323. With a sample size of 384 and P close to 0.5, the distribution of p (the estimated proportion of smokers) will be close to normal and it will be in the range of P $\pm$.05 with a chance of 19 out of 20.

The above calculations are adequate if the population size is very large. When the population size is small (say N) the standard error of p is given by

$$\sigma_p = \sqrt{\frac{N-n}{N-1}} \sqrt{\frac{PQ}{n}} \qquad (3)$$

The formula which connects the sample size to the desired degree of precision w is

$$w = z \sqrt{\frac{N-n}{N-1}} \sqrt{\frac{PQ}{n}} \qquad (4)$$

where z is the abscissa of the normal curve which cut off an area α at the tails. In the above example α was set to 1 in 20 and consequently z was 1.96. We had chosen w to be .05 in the earlier example. Equation 4 can be solved to get the required sample size for any given population size N if we are willing to make some assumption about P. As mentioned earlier the sample size will be a maximum when the value P is taken as .5 when all other factors are the same. In the absence of a firm estimate of P we will be better off to assume that P is .5 and can be sure that the error is within the desired limit.

A computer programme (FSAM) is available (from the author) to calculate the sample size for any given population size. The programme runs on any IBM compatible Personal Computer. The user can print out tables of sample sizes for six different levels of α. The programme prompts the user to select one of the six preset levels of (1-α). For any selected level of (1-α), which is referred to as "Confidence coefficient " in the print out, the programme prints out sample sizes corresponding to 6 values of w (confidence limits) varying from 0.01 to 0.10. The following is a sample output from the

programme FSAM. In most of the situations the following two tables will be adequate to decide the sample size in a survey. It can be seen that as the population size increases, the change in sample size for any combination of $(1-\alpha)$ and w is very slow. When the population size is larger than 20,000 the equation 2 can be used to calculate the sample size.

A Sample Output from the Computer Programme FSAM.

```
: This programme calculates sample sizes required :
: to obtain a given set of confidence intervals :
: and confidence coefficients for any given popula- :
: tion size. Finite population correction factor is :
: used, and this will make a difference if the popu- :
: lation size is small.                              :
:                                                    :
  You may calculate the sample size required
for a number of population sizes .
How many population sizes you would like to give?
The maximum is 20
10

Write the population sizes
 10 values should be given
200 400 800 1200 2000 4000 8000 12000 16000 20000


Write the desired confidence coefficient
associated with the sample size: select one of the
following:-
1. for 80% ; 2... 90%;  3... 95%; 4... 98%
5... 99%; 6... 99.9%
WRITE 1, 2 ..5 or 6; OR 7 TO EXIT

3
        Sample Size Required to Get the Desired Confidence
          Interval Around a Given Population Proportion

        Assumed value of the proportion       .50

            The degree of confidence      .950
```

Popln. Size	Confidence Interval (+ or -) .010	.030	.050	.070	.090	.100
200.	196	169	132	100	75	66
400.	385	292	197	132	92	78
800.	739	458	260	158	104	86
1200.	1067	566	292	169	108	89
2000.	1656	697	323	179	112	92
4000.	2825	843	351	187	116	94
8000.	4365	942	367	192	117	95
12000.	5335	981	373	193	118	96
16000.	6002	1001	376	194	118	96
20000.	6489	1014	377	195	118	96

```
Do you wish to set another level confidence cofft.?
Write "YES" OR "NO"
```

```
yes

Write the desired confidence coefficient
associated with the sample size: select one of the
following:-
1. for 80% ; 2... 90%;  3... 95%; 4... 98%
5... 99%; 6... 99.9%
WRITE 1, 2 ..5 or 6; OR 7 TO EXIT
5
```

Sample Size Required to Get the Desired Confidence Interval Around a Given Population Proportion
Assumed value of the proportion .50
The degree of confidence .990

Popln. Size	Confidence Interval (+ or -)					
	.010	.030	.050	.070	.090	.100
200.	198	181	154	126	102	91
400.	391	329	250	184	136	118
800.	764	559	363	239	164	138
1200.	1120	728	428	265	176	146
2000.	1785	960	499	290	186	154
4000.	3224	1263	570	313	195	160
8000.	5398	1499	613	325	200	163
12000.	6964	1598	629	330	202	164
16000.	8145	1653	638	332	203	165
20000.	9069	1688	643	333	203	165

```
Do you wish to set another level confidence cofft.?
Write "YES" OR "NO"
No

Do you wish to start all over again?
WRITE "YES" OR "NO"
NO
```

Sample Size when Confidence Interval for a Continuous Variable is Required.

If $\bar{y}$ is the average of the observations from a simple random sample we wish to select the sample size so that the probability of the difference between the sample mean and the unknown population mean being greater than w is no more than α, where α and w have the same meaning as in the last section. The sample mean has a normal distribution with the standard error of $\sqrt{\frac{N-n}{N-1}}\,\frac{\sigma}{\sqrt{n}}$ where σ is the unknown population standard deviation of the variable. Following the same reasoning as in the case of the categorical variables in the first part we can write,

$$w = z\sqrt{\frac{N-n}{N-1}}\,\frac{\sigma}{\sqrt{n}} \qquad (5)$$

Here z is the abscissa of the normal curve corresponding to a tail area of α. One has to have an estimate of σ to solve for the value n from equation 5. Similar studies of the variable under consideration by other researchers may give some guidance to make a guess about the amount of variation in the variable.

A computer programme SSIZE is available from the author to print out the desired sample sizes for any population size given values of w and α. This programme is more versatile than FSAM; but requires a better knowledge of the process of estimation of sample size and related theories. The user is required to give the desired widths of the confidence intervals and an estimate of the population standard deviation. The population standard deviation and w are closely related and so one should careful to give consistent values for w and the standard deviation. The following is an exmple of the output that can be obtained from this programme.

Sample Output From the Programme SSIZE

```
Write the best estimate of the standard deviation
200
You  may calculate the sample size required
for a number of population sizes.
How many population sizes you would like to give?
The maximum is 20
3

Write the population sizes
  3 values should be given
500 1000 10000

How  close to the population value would you
like the estimate from the sample to be?
In other words what should be the confidence interval?
 For example  + or - .05  .... etc
You may specify a number of confidence intervals
The maximum is 6
How many confidence intervals? (1 to 6)
3

Write the confidence intervals
  3 values are expected
10 15 20

Write the desired confidence coefficient
associated with the sample size: select one of the
following:-
1... 80%
```

```
2... 90%
3... 95%
4... 98%
5... 99%
6... 99.9% WRITE 1, 2 ..5 OR 6. PUT 7 TO EXIT

3
Sample size required to get the desired confidence
interval around a given population proportion
or population mean
Assumed value of the population standard
deviation     200.000
The degree of confidence      .950

 Population size     confidence interval (+ or -)
                  10.000  15.000  20.000

      500.           378     289     218
     1000.           607     407     278
    10000.          1333     640     370
Do you wish to set another level confidence cofft.?
Write "YES" OR "NO"
NO
```

Confidence Interval of an Estimated Proportion.

Once a sample survey has been completed the completed number of questionnires may not be the same as one had desired in the beginning. The computer programme CONFID is designed to prepare a table of confidence intervals for proportions for any combination of population size and sample size. This programme will be useful for those wish to get a quick estimate of the confidence interval where the finite population correction factor has been taken into account. A typical output from this programme is given below.

Output from the Programme CONFID

```
Write the population size and the sample size
600 100
                    Confidence intervals for diffferent
                  proportions and confidence coefficients.
                           Population size       600.
                           Sample size       100.

                              Level of confidence
p or (1-p)     .80      .90      .95      .98      .99      .999

    .05        .025     .032     .039     .046     .051     .065
    .10        .035     .045     .053     .063     .070     .090
```

	Level of confidence					
p or (1-p)	.80	.90	.95	.98	.99	.999
.15	.041	.053	.063	.075	.084	.107
.20	.046	.060	.071	.085	.094	.120
.25	.050	.065	.077	.092	.101	.130
.30	.053	.068	.082	.097	.107	.137
.35	.055	.071	.085	.101	.112	.143
.40	.057	.073	.087	.104	.115	.147
.45	.058	.074	.089	.105	.117	.149
.50	.058	.075	.089	.106	.117	.150

```
Do you wish to give another sample size
and population size?
Write "YES" OR "NO"
No
```

From the above table it can easily be seen that if the estimated proportion from a sample of 100 from a population of 600 is .45, the 95% confidence limits for the population proportion is $.45 \pm 0.05$.

References

Cochran, W. G. "Sampling Techniques". Wiley, New York , 1963.

Rossi, Peter. H., James D. Wright and Andy B. Anderson. " Handbook of Survey Research." Academic Press New York, 1983.

Sudman, S. and Bradburn. "ResponseEffects in Surveys: A review and synthesis". Aldine, Survey Research Center, Chicago.

RESEARCH DISCUSSION PAPER SERIES

69. Log-Linear Modelling and Simultaneous Standardization of Rates.
Frank Trovato. November 1990.

70. Unemployment and Occupational Aspirations: A Panel Study of High School Graduates.
Susan Empson-Warner and Harvey Krahn. November 1990.

71. An Analysis of Native Mortality in Canada.
Andrea Werner-Leonard and Frank Trovato. November 1990.

72. Ethnic Variations in Fertility: Microeconomic and Minority Group Status Effects.
Tina W.L. Chui and Frank Trovato. December 1990.

73. Quantifying Semi-Structured Interviews.
Harvey Krahn. December 1990.

74. Sampling Methods for Telephone Surveys.
N.M. Lalu. January 1991.

75. Three Models for Macro-Simulation and Population Projection.
N.M. Lalu. January 1991.

76. Migration to Quebec: A Census Analysis of British and French Canadians, 1966-71 and 1976-81.
Manoshi Das and Frank Trovato. February 1991.

77. Adolescent Smoking Intentions: Gender Differences in Peer Context.
Erica Van Roosmalen and Susan McDaniel. March 1991.

78. Socio-Cultural Mobility Issues in Canada.
Parameswara Krishnan. March 1991.

79. Political Demography of Bangladesh: A Preliminary Analysis.
A.K.M. Nurun Nabi and P. Krishnan. April 1991.

80. Implementation of CATI Techniques in an Academic Social Science Research Setting.
Dave Odynak and Cliff Kinzel. April 1991.

81. Violent and Accidental Mortality Among Immigrants in Canada.
Frank Trovato. May 1991.

82. Seat Belt Use In Edmonton: Results From a Survey.
J. Jayachandran and P. Krishnan. May 1991.

83. Structural Equation Models of Multiple Respondent Data: Mutual Influence Versus Family Factor Approaches.
M. Gillespie and J. McDonald. October 1991

84. Women and Family in the Later Years: Findings from the 1990 General Social Survey.
Susan A. McDaniel. October 1991

85. Population Projection by Single Years of Age on a Personal Computer.
N.M. Lalu. November 1991.